

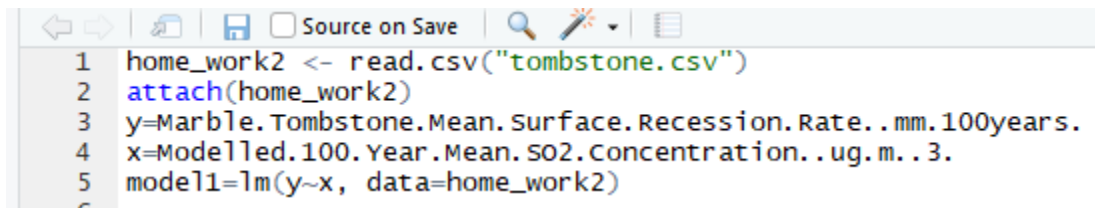
# Home Work 2

Harshith Bondada  
M10724854

**1. Read <tombstone.csv> into R. Use response variable = Marble Tombstone Mean Surface Recession Rate, and covariate = Mean SO2 concentrations over a 100 year period. Description: Marble Tombstone Mean Surface Recession Rates and Mean SO2 concentrations over a 100 year period.**

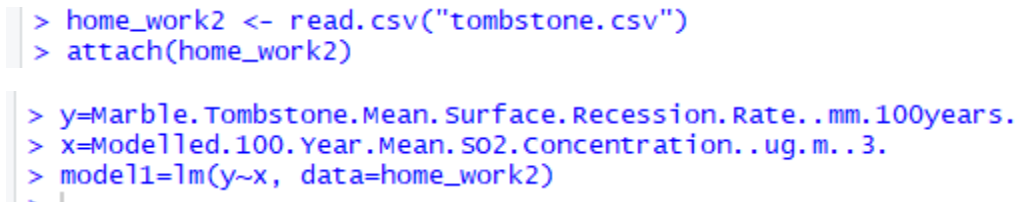
In this step, I have loaded the data into home\_work2 using the "read.csv" function.

I am taking the response variable y= Marble Tombstone surface recession rate and x= Mean SO2 concentrations. I have generated the model1 for the data.



```
1 home_work2 <- read.csv("tombstone.csv")
2 attach(home_work2)
3 y=Marble.Tombstone.Mean.Surface.Recession.Rate..mm.100years.
4 x=Modelled.100.Year.Mean.SO2.Concentration..ug.m..3.
5 model1=lm(y~x, data=home_work2)
```

This snippet shows that the commands are successfully executed. And model1 has been generated.

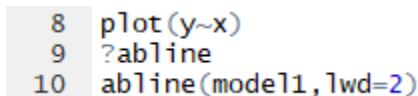


```
> home_work2 <- read.csv("tombstone.csv")
> attach(home_work2)

> y=Marble.Tombstone.Mean.Surface.Recession.Rate..mm.100years.
> x=Modelled.100.Year.Mean.SO2.Concentration..ug.m..3.
> model1=lm(y~x, data=home_work2)
```

**2. Plot data and briefly describe what you observe.**

I have plotted the data using the plot command.



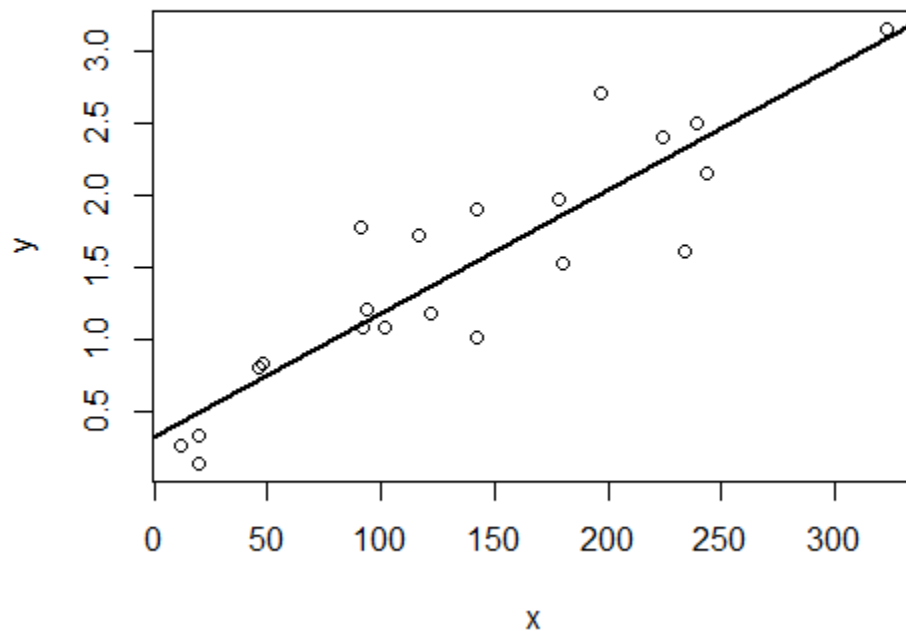
```
8 plot(y~x)
9 ?abline
10 abline(model1,lwd=2)
```

# Home Work 2

Harshith Bondada  
M10724854

The below image shows that there is a positive trend in the regression line.

Low x-values correspond to low y-values, and high x-values correspond to high y-values



### 3. Perform linear regression using lm() function

#### 3.1. Obtain coefficient estimates $\hat{\beta}_0, \hat{\beta}_1$ .

The coefficient estimates can be obtained using the `model1$coefficients` command.

```
21 model1$coefficients
```

We get the  $B_0^{\wedge}$  and  $B_1^{\wedge}$ .

```
> model1$coefficients
(Intercept)          x
0.322995899 0.008593333
```

# Home Work 2

Harshith Bondada  
M10724854

## 3.2. Obtain fitted values and the sum of fitted values.

We get the fitted values and their sum by using the following command.

model1\$fitted.values for generating the fitted values and for finding the sum we can simply get the sum using the "sum" function.

```
14 model1$fitted.values
15 sum(model1$fitted.values)
```

This below screenshot shows the output of the fitted values and their sum.

```
> model1$fitted.values
      1      2      3      4      5      6      7      8      9
0.4261159 0.4948626 0.4948626 0.7182892 0.7354759 1.1135825 1.1049892 1.1307692 1.1995159
     10     11     12     13     14     15     16     17     18
1.3284159 1.3713825 1.5432492 1.5432492 1.8526092 1.8697959 2.0158825 2.2479025 2.3338359
     19     20     21
2.3768025 2.4197692 3.0986425
> sum(model1$fitted.values)
[1] 31.42
```

## 3.3. Obtain the sum of all values of response variable.

For calculating the sum of the response variable we simply use the sum function.

Since, Y represents the response variable, we simply calculate the sum of Y.

```
8 sum(y)
```

The sum of the response variable is shown below.

```
> sum(y)
[1] 31.42
```

## 3.4. Obtain residuals and the sum of residuals.

We get the residuals and it's sum using the commands which we used for the fitted values earlier. Except, we here use the residuals.

# Home Work 2

Harshith Bondada  
M10724854

```
17 model1$residuals
18 sum(model1$residuals)
```

The residuals and their sum is:

```
> model1$residuals
      1      2      3      4      5      6
-0.15611590 -0.35486256 -0.16486256  0.09171078  0.10452411 -0.03358255
      7      8      9     10     11     12
 0.67501079  0.07923079 -0.10951588  0.39158412 -0.19138254 -0.53324921
     13     14     15     16     17     18
 0.35675079  0.12739080 -0.33979586  0.69411747  0.16209748 -0.72383585
     19     20     21
 0.13319748 -0.26976919  0.06135750
> |
> sum(model1$residuals)
[1] -3.191891e-16
```

## 3.5. Obtain the standard errors of $\hat{\beta}_0, \hat{\beta}_1$ .

For calculating the standard errors we use the following function called "summary".

This bring the entire details for our generated models.

The second line of the command returns the second column of the summary which is the standard error.

```
17 summary(model1)$coef
18 summary(model1)$coef[,2]
```

This is the summary. We can see that the standard error of the intercept and X are shown below.

```
> summary(model1)$coef
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.322995899 0.1521958377  2.122239 4.718525e-02
x            0.008593333 0.0009499341  9.046242 2.578534e-08
> summary(model1)$coef[,2]
(Intercept)      x
0.1521958377 0.0009499341
.
```

# Home Work 2

Harshith Bondada

M10724854

## 4. Suppose we increase SO2 Concentration by one unit, how does such a change influence the Surface Recession Rate?

If we increase the SO2 concentration by one unit we can see that the  $B_0$   $B_1$  change and we observe that there is a slight change in the plot. We can see that the residual values change very slightly.

We also see that the standard error for the Y intercept has increased slightly.

SE for Y intercept in model1 was around 0.152 whereas, in model2 it was 0.153

We can also see that the P value is slightly changed.

Therefore, We can say that the increase in concentration by one unit doesn't have that much change.

You can find the summary for both the models in the screenshot below.

```
> summary(model1)
```

```
Call:
```

```
lm(formula = y ~ x, data = home_work2)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-0.72384 -0.19138  0.06136  0.13320  0.69412
```

```
Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3229959   0.1521958    2.122   0.0472 *
x             0.0085933   0.0009499    9.046 2.58e-08 ***
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.365 on 19 degrees of freedom
```

```
Multiple R-squared:  0.8116,    Adjusted R-squared:  0.8017
```

```
F-statistic: 81.83 on 1 and 19 DF,  p-value: 2.579e-08
```

# Home Work 2

Harshith Bondada  
M10724854

```
> summary(model2)

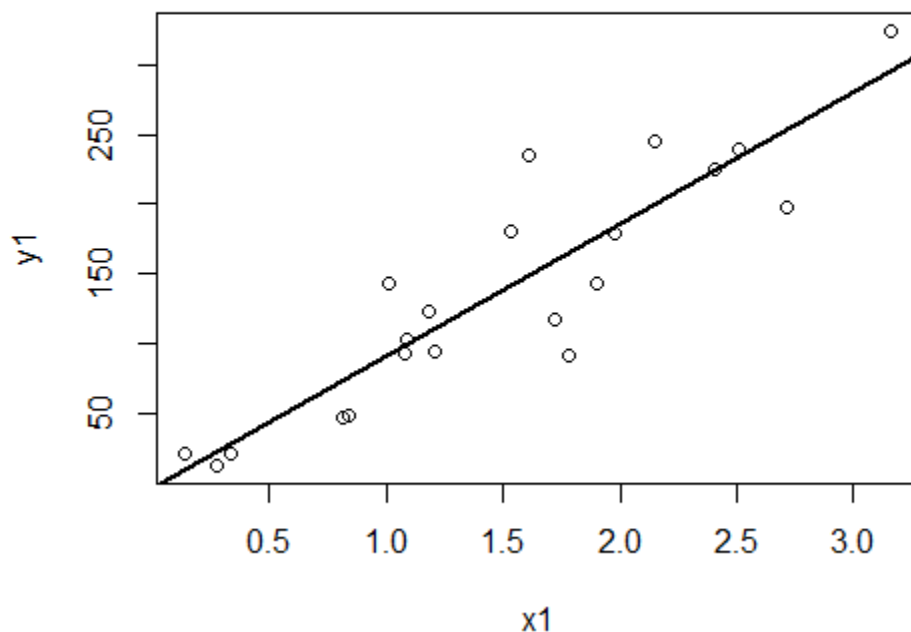
call:
lm(formula = y1 ~ x1, data = work)

Residuals:
    Min       1Q   Median       3Q      Max
-0.72384 -0.19138  0.06136  0.13320  0.69412

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3144026   0.1530061    2.055   0.0539 .
x1           0.0085933   0.0009499    9.046 2.58e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.365 on 19 degrees of freedom
Multiple R-squared:  0.8116,    Adjusted R-squared:  0.8017
F-statistic: 81.83 on 1 and 19 DF,  p-value: 2.579e-08
```

We can find in the below plot that the residual values almost have no change.



# Home Work 2

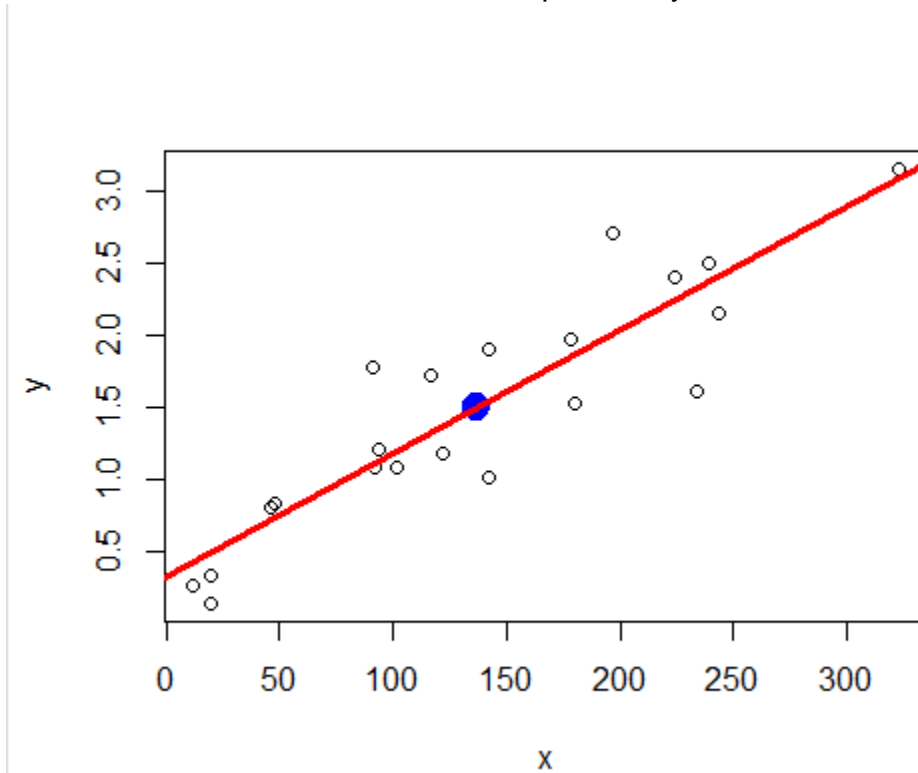
Harshith Bondada  
M10724854

**5. Does the intercept of the linear regression have natural interpretation? If so, what does it mean?**

Yes, the intercept of the linear regression has natural interpretation.

It says, that the intercept is well within the data range and the intercept is on the positive y-axis which means that the linear regression intercept is valid.

If we follow the red line it intercepts the y axis well within the positive range.



```
> model1$coefficients
(Intercept)          x
0.322995899 0.008593333
```

For example, when we take the value of X to be 50, when we calculate the with the  $B_0^{\wedge}$  and  $B_1^{\wedge}$  we get the y intercept to be around 0.7 which is correct according to the graph.

```
> sum(model1$fitted.values)
[1] 31.42
> sum(y)
[1] 31.42
```

# Home Work 2

Harshith Bondada

M10724854

We can also find that the sum of the observed values is equal to the sum of the fitted values. This means that the regression line is valid and is in natural setting.

## 6. Which area (i.e., observation) has the highest Surface Recession Rate?

I wrote the code for calculating which observation has the highest Surface Recession Rate.

I just stored the first value of the response variable in S.

I then compared it with all the values which in turn returns the value with the highest recession rate.

```
45 Largest_area=City[1]
46 s=y[1]
47 for(j in 1:length(y))
48 {
49     if(y[j]>s)
50     {
51         Largest_area=City[j]
52     }
53 }
54 sprintf("The largest area with highest surface recession rate is %s", Largest_area)
```

The area with the highest recession rate is displayed in the form of the message.

I used the sprintf function for printing the message.

```
> sprintf("The largest area with highest surface recession rate is %s", Largest_area)
[1] "The largest area with highest surface recession rate is Chicago,IL"
```

## 7. Which area (i.e., observation) has the largest residual (i.e., the largest absolute value) according to the linear regression you just fitted?

In the I am displaying the observation number as well as the observation name with the highest residual value.

First, I have calculated the which is the maximum in the residual value using the MAX function. And then using the which command returned the observation index.

```
38 b=which(model1$residuals==max(model1$residuals))
39 b
40 sprintf("The observation which has the highest residual value is:%s", city[b])
41 city[b]
```

This screenshot shows the observation index which is 16 and then displaying the name with the highest residual value which is Pittsburgh, PA.



## Home Work 2

Harshith Bondada  
M10724854

```
[1] The observation which has the highest residual value is: Pittsburgh, PA
> b=which(model1$residuals==max(model1$residuals))
> b
[1] 16
> sprintf("The observation which has the highest residual value is:%s", City[b])
[1] "The observation which has the highest residual value is:Pittsburgh,PA"
```

**8. Calculate the mean of covariate and mean of response. Verify the fact that the fitted regression line go through the point  $(\bar{x}, \bar{y})$ .**

First, calculate the mean for the response variable and the covariate.

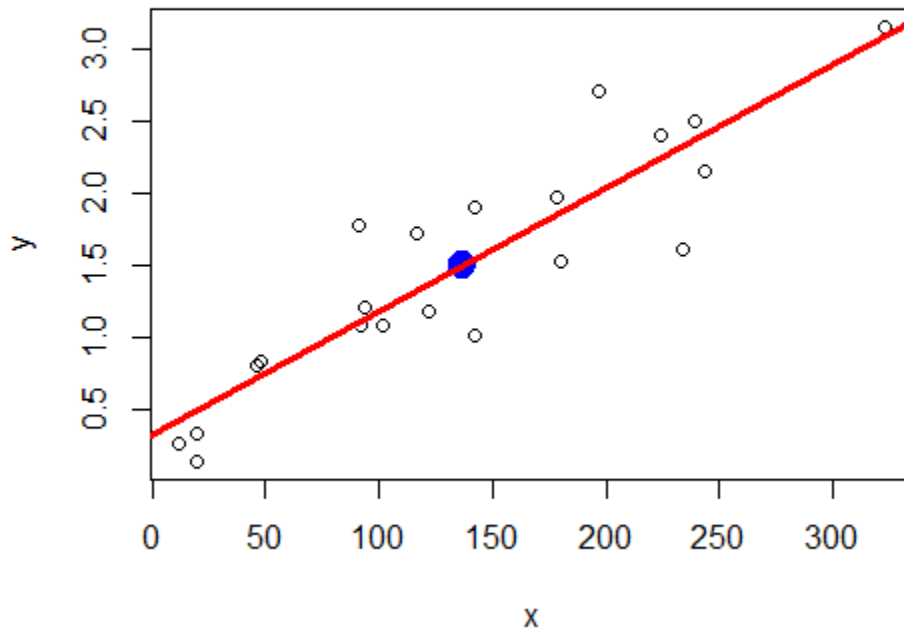
Then plot the point for the mean. Plot the regression line using the abline function and we see that it passes through the mean point.

```
40 mean(y)
41 mean(x)
42 points(mean(x),mean(y),pch=20,col="blue",cex=3)
43 abline(lm(y~x),col="red",lwd=3)
```

The below is the graph which shows that the regression line passes through the mean point.

## Home Work 2

Harshith Bondada  
M10724854



This below screenshot shows the mean of the response variable and the covariate.

```
> mean(y)
[1] 1.49619
> mean(x)
[1] 136.5238
> points(mean(x),mean(y),pch=20,col="blue",cex=3)
> abline(lm(y~x),col="blue",lwd=3)
> abline(lm(y~x),col="red",lwd=3)
```

# Home Work 2

Harshith Bondada

M10724854

9. Repeat the same questions (1-8) for the data set <bus.csv>. Description: Cross-sectional analysis of 24 British bus companies (1951). Use response variable = Expenses per car mile (pence), covariate = Car miles per year (1000s).

In this step, I have loaded the data into Home\_Work using the "read.csv" function.

I am taking the response variable y = Expenses per car mile

and x = Car miles per year. I have generated the model3 for the data.

```
1 Home_work <- read.csv("bus.csv")
2 attach(Home_work)
3 y2=Expenses.per.car.mile..pence.
4 x2=Car.miles.per.year..1000s.
5 model3=lm(y~x, data=Home_work)
```

This snippet shows that the commands are successfully executed. And model1 has been generated.

```
> Home_work <- read.csv("bus.csv")
> attach(Home_work2)

> y2=Expenses.per.car.mile..pence.
> x2=Car.miles.per.year..1000s.
> model3=lm(y~x, data=Home_work)
> summary(model1)

Call:
```

9.2. Plot data and briefly describe what you observe.

I have plotted the data using the plot command.

```
10 plot(y2~x2)
11 ?abline
12 abline(model3,lwd=2)
13
```

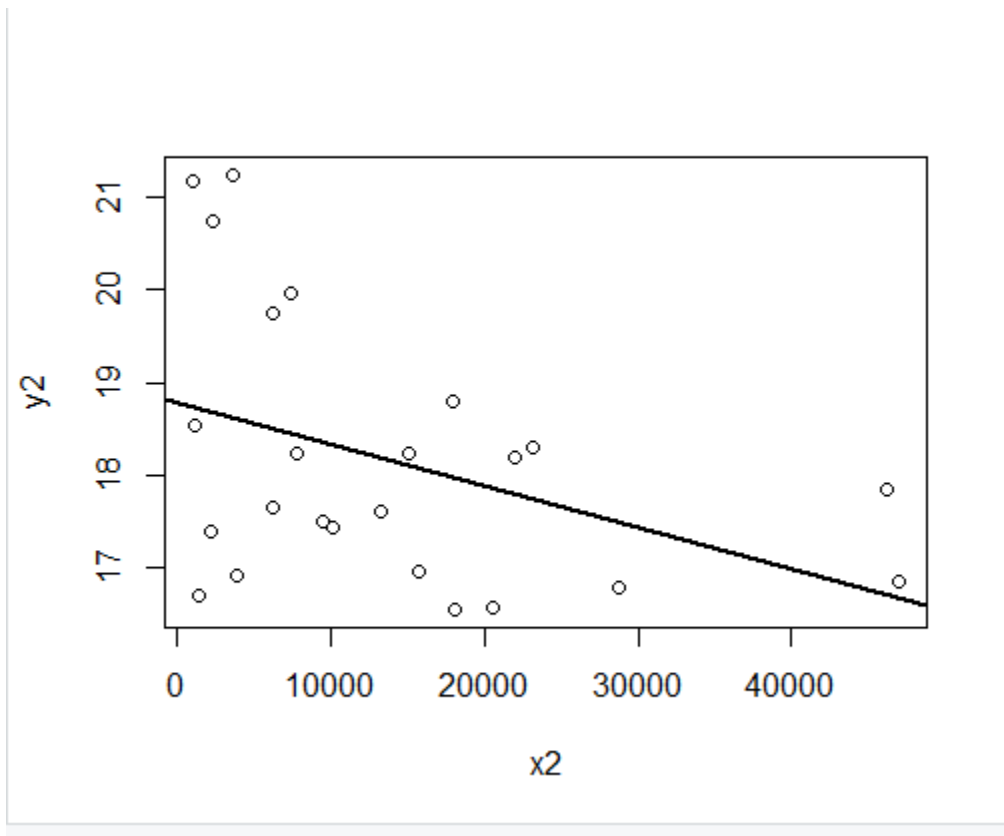
The below image shows that there is a negative trend in the regression line.

# Home Work 2

Harshith Bondada

M10724854

When the slope of the regression line is negative (meaning that the value of  $b$  is negative) the value of  $y$  decreases as  $x$  increases



## 9.3. Perform linear regression using `lm()` function

### 9.3.1. Obtain coefficient estimates $\hat{\beta}_0, \hat{\beta}_1$ .

The coefficient estimates can be obtained using the `model1$coefficients` command.

```
21 model1$coefficients
```

We get the  $B_0^{\wedge}$  and  $B_1^{\wedge}$ .

# Home Work 2

Harshith Bondada  
M10724854

```
> model3$coefficients
      (Intercept)          x2
1.878180e+01 -4.449914e-05
```

## 9.3.2. Obtain fitted values and the sum of fitted values.

We get the fitted values and their sum by using the following command.

model3\$fitted.values for generating the fitted values and for finding the sum we can simply get the sum using the "sum" function.

```
19 model3$fitted.values
20 sum(model3$fitted.values)
```

The below screenshot shows the output for fitted values and their sum.

```
> model3$fitted.values
      1      2      3      4      5      6      7      8
18.50435 16.72461 18.45429 17.50401 17.80576 18.72231 17.98611 18.67861
      9     10     11     12     13     14     15     16
17.97904 18.73076 18.68497 18.19143 18.62245 18.10969 16.68994 18.33063
     17     18     19     20     21     22     23     24
18.50827 17.75436 17.86734 18.36129 18.73606 18.61057 18.08512 18.43805
> sum(model3$fitted.values)
[1] 436.08
```

## 9.3.3. Obtain the sum of all values of response variable.

For calculating the sum of the response variable we simply use the sum function.

Since, Y represents the response variable, we simply calculate the sum of Y.

```
8 sum(y2)
```

The sum of the response variable is shown below:

```
> sum(y2)
[1] 436.08
>
```

## 9.3.4. Obtain residuals and the sum of residuals.

# Home Work 2

Harshith Bondada

M10724854

We get the residuals and it's sum using the commands which we used for the fitted values earlier. Except, we here use the residuals.

```
21 model3$residuals
22 sum(model3$residuals)
```

The residuals and their sum is shown in the below screenshot

```
> model3$residuals
      1      2      3      4      5      6      7
1.2556501 1.1253933 1.5057117 -0.7040092 0.3942422 -2.0123067 0.8238871
      8      9     10     11     12     13     14
2.0613915 -1.4190375 -0.1807615 -1.2849719 -0.5714319 2.6175494 0.1203130
     15     16     17     18     19     20     21
0.1700581 -0.8806252 -0.8482658 0.5456387 -1.2873447 -0.8512851 2.4339431
     22     23     24
-1.6905693 -1.1251235 -0.1980461
> sum(model3$residuals)
[1] 1.637579e-15
```

## 9.3.5. Obtain the standard errors of $\hat{\beta}_0, \hat{\beta}_1$ .

For calculating the standard errors we use the following function called "summary".

This bring the entire details for our generated models.

The second line of the command returns the second column of the summary which is the standard error.

```
25 summary(model3)$coef|
26 summary(model3)$coef[,2]
```

This is the summary. We can see that the standard error of the intercept and X are shown below.

```
> summary(model3)$coef
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.878180e+01 4.075464e-01 46.08506 2.223005e-23
x2           -4.449914e-05 2.187948e-05 -2.03383 5.420264e-02
> summary(model3)$coef[,2]
      (Intercept)      x2
4.075464e-01 2.187948e-05
```

## Home Work 2

Harshith Bondada

M10724854

9.4. Suppose we increase car miles by one unit, how does such a change influence the expenses?

If we increase the car miles by one unit we can see that the  $B^0$   $B^1$  doesn't change and we observe that there is a slight change in the plot. We can see that the residual values are also not changing.

We also see that the standard error for the Y intercept has neither increased or decreased.

Therefore, We can say that the increase in concentration by one unit doesn't have that much change.

You can find the summary for both the models in the screenshot below.

```
> summary(model3)

Call:
lm(formula = y2 ~ x2, data = Home_work)

Residuals:
    Min       1Q   Median       3Q      Max
-2.0123 -0.9417 -0.1894  0.8993  2.6176

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.878e+01  4.075e-01  46.085  <2e-16 ***
x2           -4.450e-05  2.188e-05  -2.034   0.0542 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.347 on 22 degrees of freedom
Multiple R-squared:  0.1583,    Adjusted R-squared:  0.12
F-statistic: 4.136 on 1 and 22 DF,  p-value: 0.0542
```

# Home Work 2

Harshith Bondada  
M10724854

```
> summary(model4)
```

call:

```
lm(formula = y3 ~ x3, data = Home_work2)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.0123	-0.9417	-0.1894	0.8993	2.6176

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.878e+01	4.076e-01	46.083	<2e-16 ***
x3	-4.450e-05	2.188e-05	-2.034	0.0542 .

---

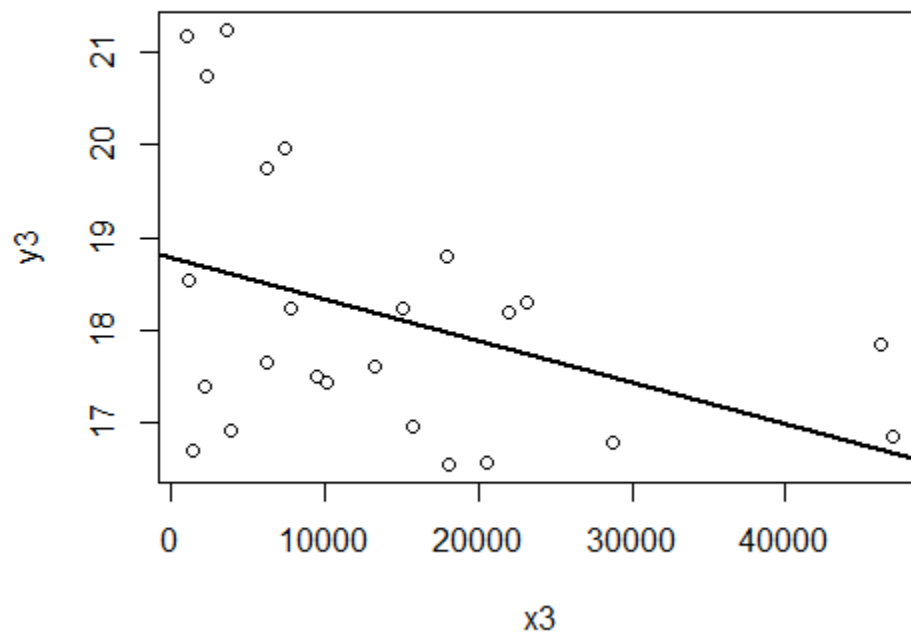
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.347 on 22 degrees of freedom

Multiple R-squared: 0.1583, Adjusted R-squared: 0.12

F-statistic: 4.136 on 1 and 22 DF, p-value: 0.0542

We can find in the below plot that the residual values almost have no change.





# Home Work 2

Harshith Bondada

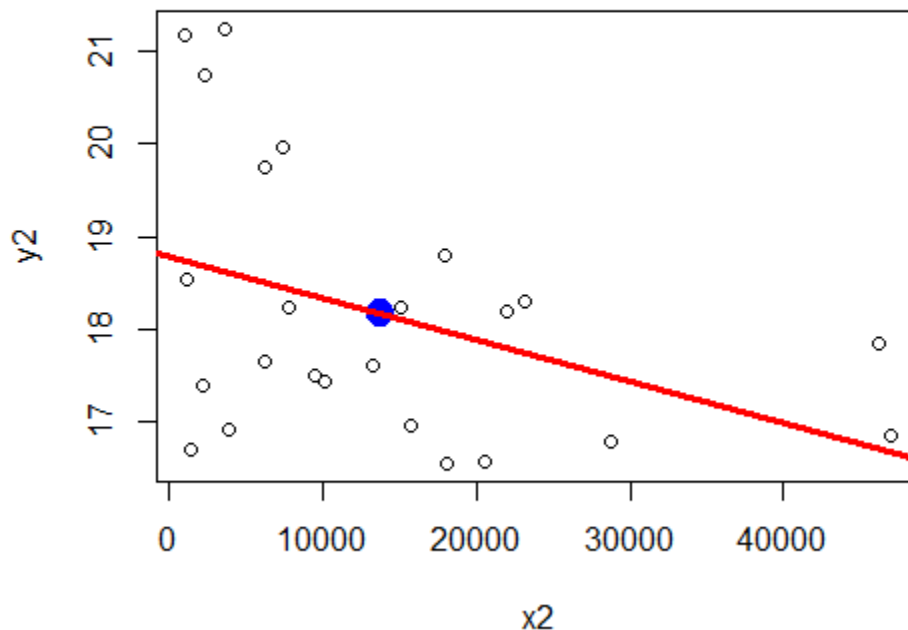
M10724854

**9.5. Does the intercept of the linear regression have natural interpretation? If so, what does it mean?**

Yes, the intercept of the linear regression has natural interpretation.

When the slope of the regression line is negative (meaning that the value of  $b$  is negative) the value of  $y$  decreases as  $x$  increases.

We can see in the figure that the value of  $y$  decreases as the value of  $x$  increases.



```
> model3$fitted.values
 1      2      3      4      5      6      7      8
18.50435 16.72461 18.45429 17.50401 17.80576 18.72231 17.98611 18.67861
 9     10     11     12     13     14     15     16
17.97904 18.73076 18.68497 18.19143 18.62245 18.10969 16.68994 18.33063
17     18     19     20     21     22     23     24
18.50827 17.75436 17.86734 18.36129 18.73606 18.61057 18.08512 18.43805
> sum(model3$fitted.values)
[1] 436.08

> sum(y2)
[1] 436.08
> |
```

We can also find that the sum of the observed values is equal to the sum of the fitted values. This means that the regression line is valid and is in natural setting

# Home Work 2

Harshith Bondada  
M10724854

```
> model3$coefficients
      (Intercept)          x2
1.878180e+01 -4.449914e-05
```

For example, when we take the value of X to be 10000, when we calculate the with the  $B_0^{\wedge}$  and  $B_1^{\wedge}$  we get the y intercept to be around 17.5 which is correct according to the graph.

## 9.6. Which area (i.e., observation) has the highest Expenses per car mile?

For this question, I am using the max function to get the highest observation value of the response variable and then using the which function to get the index of the highest observation.

```
26 which(Home_work$Expenses.per.car.mile..pence.==max(y2))
```

The below screenshot shows the highest observation index:

```
> which(Home_work$Expenses.per.car.mile..pence.==max(y2))
[1] 13
```

## 9.7. Which area (i.e., observation) has the largest residual (i.e., the largest absolute value) according to the linear regression you just fitted?

In the I am displaying the observation number as well as the observation name with the highest residual value.

First, I have calculated the which is the maximum in the residual value using the MAX function. And then using the which command returned the observation index.

```
26 a=which(model3$residuals==max(model3$residuals))
27 a
```

The observation with the highest residual value is shown below:

```
> a=which(model3$residuals==max(model3$residuals))
> a
13
13
```

## Home Work 2

Harshith Bondada  
M10724854

**9.8. Calculate the mean of covariate and mean of response. Verify the fact that the fitted regression line go through the point  $(\bar{x}, \bar{y})$ .**

First, calculate the mean for the response variable and the covariate.

Then plot the point for the mean. Plot the regression line using the abline function and we see that it passes through the mean point.

```
53 mean(y2)
54 mean(x2)
55 points(mean(x2),mean(y2),pch=20,col="blue",cex=3)
56 abline(lm(y2~x2),col="red",lwd=3)
```

The below is the graph which shows that the regression line passes through the mean point.

